Dependent Hierarchical Normalized Random Measures for Dynamic Topic Modeling

Changyou Chen[1,2], Ding Nan[3], Wray Buntine[2,1]
[1]Australian National University; [2]National ICT, Canberra, ACT, Australia; [3]Purdue University
Changyou.Chen@NICTA.com.au, ding10@purdue.edu, Wray. Buntine@NICTA.com.au

1 Motivation

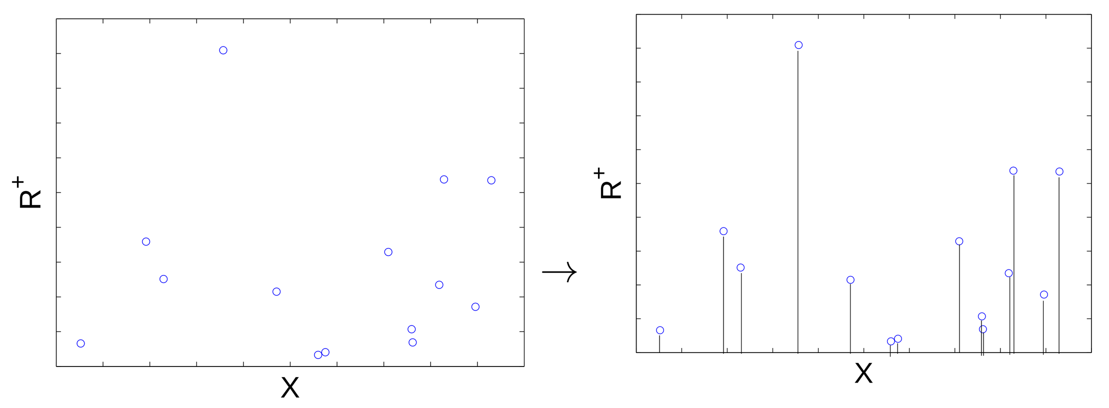
- We want to model the birth-death process of topic evolution.
- We want to model the topic dependency between time frames.
- We want to model the power-law phenomena appeared in most of natural datasets, *e.g.*, text datasets.

2 Normalized Random Measures

Poisson Processes: A *Poisson process* on $\mathbb{S}$ is a random subset $\Pi \in \mathbb{S}$ such that if $N(A)$ is the number of points of Π in $A \subseteq \mathbb{S}$, then $N(A)$ is a Poisson random variable with mean $\nu(A)$, and $N(A_1), \cdots, N(A_n)$ are independent if $A_1, \cdots, A_n$ are disjoint.

Completely Random Measures (CRM): Let $\mathbb{S} = R^+ \times \mathbb{X}$, a CRM $\tilde{\mu}$ is defined as a linear functional of the Poisson random measure $N(\cdot)$ (called $\nu(\cdot)$ the Lévy measure of $\tilde{\mu}$)

$$\tilde{\mu}(B) = \int_{\mathbb{R}^+ \times B} t N(\mathrm{d}t, \mathrm{d}x), \forall B \in \mathcal{B}(\mathbb{X}).$$



Poisson processes: $N(A) = \sum_{(J(x),x) \in A} \delta_{(J(x),x)}$ **Completely random measures:** $\tilde{\mu}(A) = \sum_{(J(x),x) \in A} J(x)\delta_x$

Normalized Random Measures (NRM): An NRM is obtained by normalizing the CRM $\tilde{\mu}$ as: $\mu = \frac{\tilde{\mu}}{\tilde{\mu}(\mathbb{X})}$. A normalized generalized Gamma process (NGG) is an NRM with Lévy measure being $\frac{e^{-bt}}{t^{1+a}} H(\mathrm{d}x), b > 0, 0 < a < 1$.

Normalized Generalized Gamma Process (NGG): A normalized generalized Gamma process (NGG) is an NRM with Lévy measure being $\frac{e^{-bt}}{t^{1+a}} H(\mathrm{d}x)$, where $0 < a < 1, b > 0$.

3 The three Dependency Operations

Superposition of NRMs: Given n independent NRMs $\mu_1, \cdots, \mu_n$ on $\mathbb{X}$, the superposition ($\oplus$) is:

$$\mu_1 \oplus \mu_2 \oplus \cdots \oplus \mu_n := c_1\mu_1 + c_2\mu_2 + \cdots + c_n\mu_n .$$

where the weights $c_m = \frac{\tilde{\mu}_m(\mathbb{X})}{\sum_j \tilde{\mu}_j(\mathbb{X})}$ and $\tilde{\mu}_m$ is the unnormalized random measures corresponding to μ_m.

Subsampling of NRMs: Given a NRM $\mu = \sum_{k=1}^{\infty} r_k \delta_{\theta_k}$ on $\mathbb{X}$, and a Bernoulli parameter $q \in [0, 1]$, the subsampling of μ, is defined as

$$S^q(\mu) := \sum_{k:z_k=1} \frac{r_k}{\sum_j z_j r_j} \delta_{\theta_k},$$

where $z_k \sim$ Bernoulli(q) are Bernoulli random variables with acceptance rate q.

Point transition of NRMs: Given a NRM $\mu = \sum_{k=1}^{\infty} r_k \delta_{\theta_k}$ on $\mathbb{X}$, the point transition of μ, is to draw atoms θ'_k from a transformed base measure to yield a new NRM as

$$T(\mu) := \sum_{k=1}^{\infty} r_k \delta_{\theta'_k} .$$

4 Sampling

The statistics we are interested in are:

- x_{mji}: the customer i in the jth restaurant.
- s_{mji}: the dish that x_{mji} is eating.
- n'_{mk}: $n'_{mk} = \sum_j \sum_r \delta_{\psi_{mjr}=k}$, the number of customers in μ'_m eating dish k.
- $\tilde{\mu}_m = \sum_k J_{mk}\delta_{\theta_k}$, $\tilde{\mu}'_m = \sum_k J'_{mk}\delta_{\theta_k}$.

At each time frame m, we do:

- Slice sample J_{mk} (ends up finite jumps).
- Subsample J'_{mk} by inheriting from $J_{m'k}, m' \le m$ with Bernoulli trials.
- Construct μ'_m by normalizing J'_{mk}.
- Sample s_{mji} using a generalized Blackwell-MacQueen sampling scheme for the hierarchical NRM.
- Sample n'_{mk} by simulating a generalized Chinese restaurant process for the NRM.

5 Experiments

Evaluated on 9 datasets including *news, blogs, academic* and *Twitter* collections. See Figure 1, 2, 3 for demonstration and Table 1 for comparison.

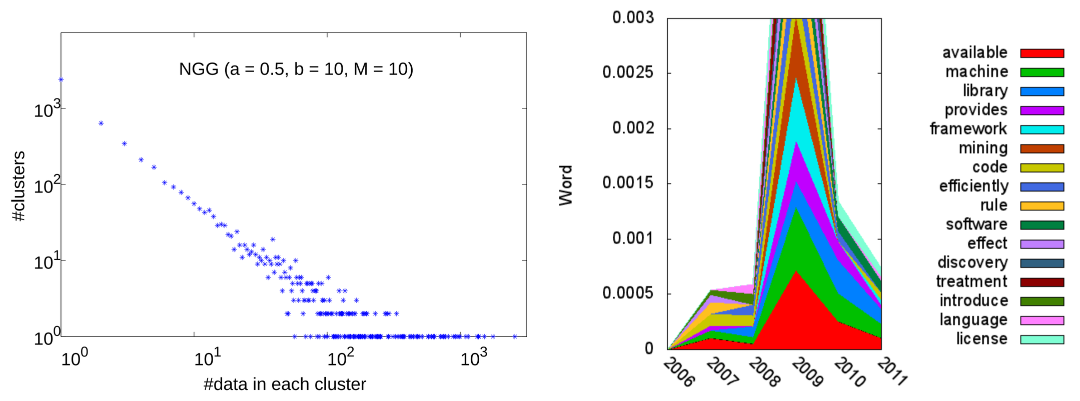


Figure 1: Left: Power-law phenomena in NGG; Right: topic evolution on JMLR. Shows a late developing topic on software, before during and after the start of MLOSS.org in 2008.

Table 1: Test log-likelihood on 9 datasets. *DHNGG*: dependent hierarchical normalized generalized Gamma processes, *DHDP*: dependent hierarchical Dirichlet processes, *HDP*: hierarchical Dirichlet processes, *DTM:* dynamic topic model.

Datasets	ICML	JMLR	TPAMI	NIPS	Person
DHNGG	**-5.3123e+04**	**-7.3318e+04**	**-1.1841e+05**	**-4.1866e+06**	**-2.4718e+06**
DHDP	-5.3366e+04	-7.3661e+04	-1.2006e+05	-4.4055e+06	-2.4763e+06
HDP	-5.4793e+04	-7.7442e+04	-1.2363e+05	-4.4122e+06	-2.6125e+06
DTM	-6.2982e+04	-8.7226e+04	-1.4021e+05	-5.1590e+06	-2.9023e+06
Datasets	Twitter$_1$	Twitter$_2$	Twitter$_3$	BDT	
DHNGG	**-1.0391e+05**	**-2.1777e+05**	**-1.5694e+05**	**-3.3909e+05**	
DHDP	-1.0711e+05	-2.2090e+05	-1.5847e+05	-3.4048e+05	
HDP	-1.0752e+05	-2.1903e+05	-1.6016e+05	-3.4833e+05	
DTM	-1.2130e+05	-2.6264e+05	-1.9929e+05	-3.9316e+05	

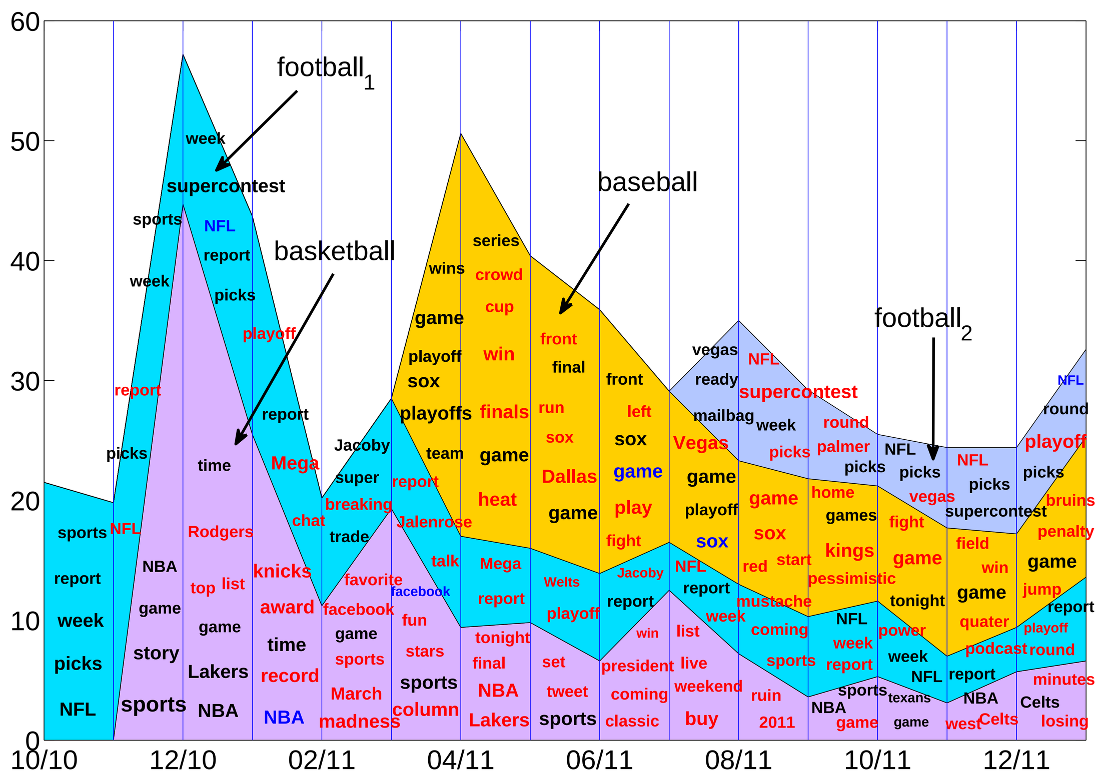


Figure 2: Topic evolution on Twitter. Words in red have increased, and blue decreased.

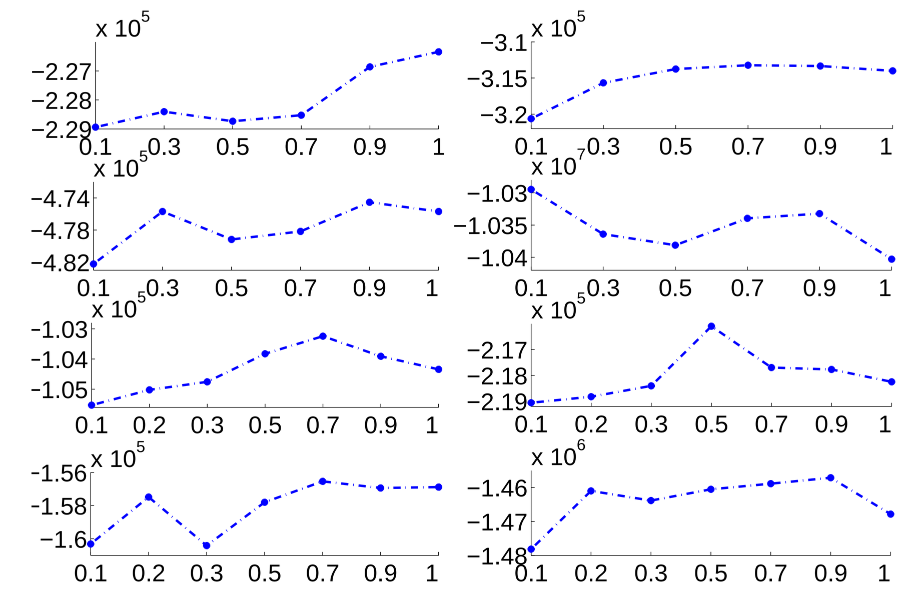


Figure 3: Training log-likelihoods influenced by the subsampling rate q. From top-down, left to right are the results on ICML, JMLR, TPAMI, Person, Twitter$_1$, Twitter$_2$, Twitter$_3$ and BDT datasets, respectively.

Theorem 1 *The time dependent random measures represented in Figure 4 are equivalent. Furthermore, both resulting NRMs μ'_m's are equal to:*

$$\mu'_m = \sum_{j=1}^{m} \frac{(q^{m-j}\tilde{\mu}_j)(\mathbb{X})}{\sum_{j'=1}^{m} (q^{m-j'}\tilde{\mu}_{j'})(\mathbb{X})} T_{m-j}(\mu_j), m > 1$$

where $q^{m-j}\tilde{\mu}$ is the random measure with Lévy measure $q^{m-j}\nu(\mathrm{d}t, \mathrm{d}x)$ ($\nu(\mathrm{d}t, \mathrm{d}x)$ is the Lévy measure of $\tilde{\mu}$). $T_{m-j}(\mu)$ denotes point transition on μ for $(m-j)$ times .

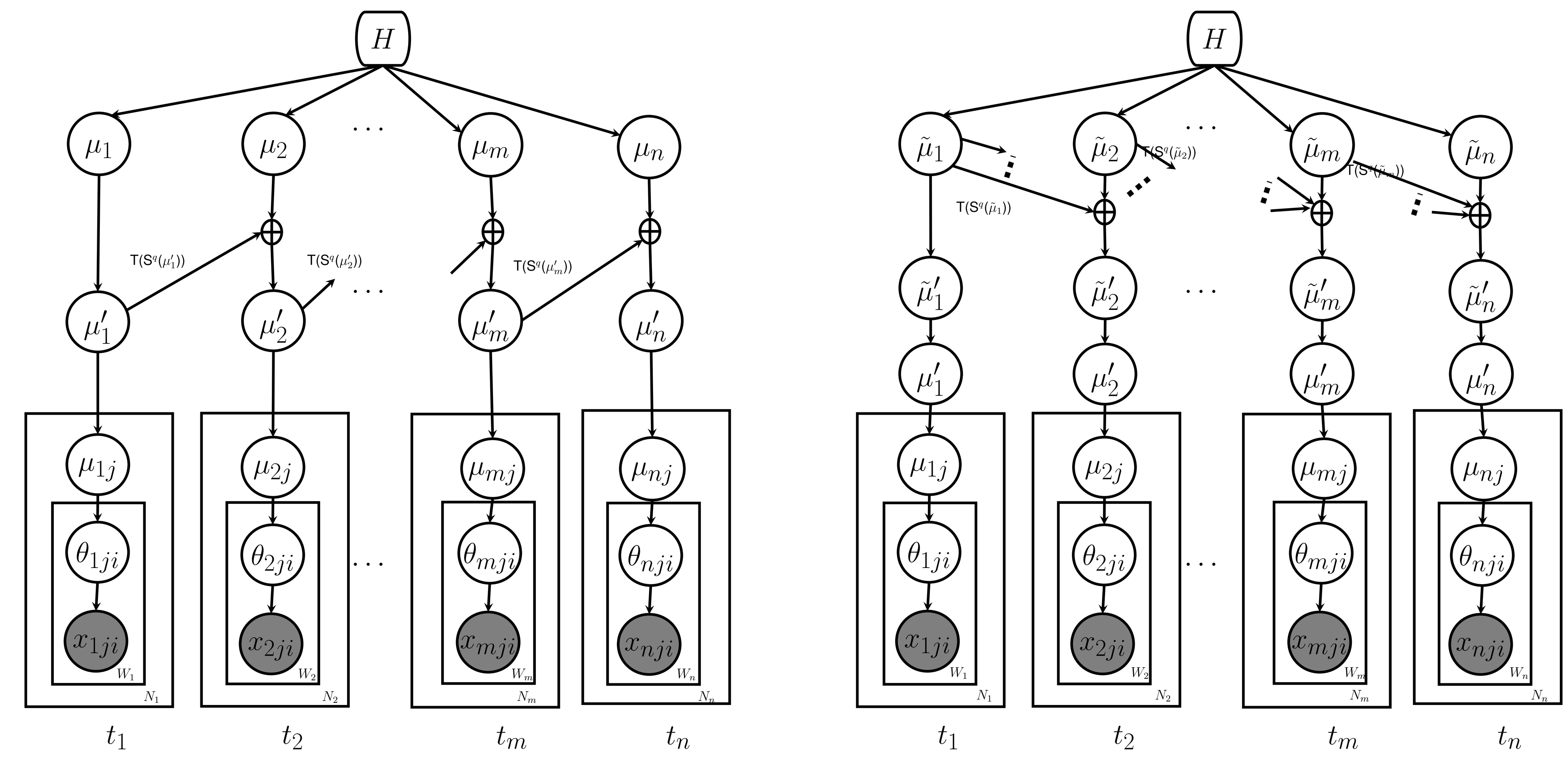


Figure 4: The time dependent topic model. The left plot corresponds to directly manipulating on normalized random measures, the right one corresponds to manipulating on completely random measures. T: Point transition; S^q: Subsampling with acceptance rate q; $\oplus$: Superposition. Here $m = n - 1$ in the figures.

Generative Process:

- Generating independent NRMs μ_m for time frame $m = 1, \cdots, n$:

$$\mu_m | H, \eta_0 \sim \text{NRM}(M_0, \eta_0, P_0) \quad (1)$$

where $H(\cdot) = M_0 P_0(\cdot)$. M_0 is the total mass for μ_m and P_0 is the base distribution. η_0 is the set of hyperparameters of the corresponding NRM.

- Generating dependent NRMs μ'_m (from μ_m and μ'_{m-1}), for time frame $m > 1$:

$$\mu'_m = T(S^q(\mu'_{m-1})) \oplus \mu_m . \quad (2)$$

- Generating hierarchical NRM mixtures ($\mu_{mj}, \theta_{mji}, x_{mji}$) for time frame $m = 1, \cdots, n$, document $j = 1, \cdots, N_m$, word $i = 1, \cdots, W_{mj}$:

$$\mu_{mj} = \text{NRM}(M_m, \eta_m, \mu'_m), \quad (3)$$
$$\theta_{mji} | \mu_{mj} \sim \mu_{mj}, \quad x_{mji} | \theta_{mji} \sim g_0(\cdot | \theta_{mji})$$

where M_m is the total mass for μ_{mj}, $g_0(\cdot|\theta_{mji})$ denotes the density function to generate data x_{mji} from atom θ_{mji}.

Statistical Machine Learning (SML) Group
ANU College of Engineering & Computer Science
Australian National University
NICTA